Wide Area Information Servers: A Supercomputer on every Desk

Brewster Kahle
Thinking Machines Corporation

What I really want...

- My personal information to be accessible
- Published information should find me
- Usable anywhere
- Others can use what I have learned (if I want them to)

What is it?

Electronic Publishing

(Or publishing over wires)

New Communications Technology Problems

	BOOKS	
Experts only	*Monks*	
Distribution is hard and expensive	*Vellum is calf skin*	
Different interfaces	*1000's of languages in Europe alone*	
Material is intractable	*Scrolls and manuscripts were about as random access as musical scores*	
Business model needed	*Centralized printing*	

Telegraph>
Telephone

Operators

Telephones on barb wire

Switching was manual

No white pages

Pay per minute

Electronic Publishing

Professional searchers

$1/minute over obscure modems

//query (W5) inform?

600 databases on Dialog
~1 Terabyte
140Gbyte at DJ
80GB card catalog at RLG

Not understood

New Communications Technology Problems

	BOOKS	Telegraph> Telephone	Electronic Publishing
Experts only	*Monks*	*Operators*	*Professional searchers*
Distribution is hard and expensive	*Vellum is calf skin*	*Telephones on barb wire*	*$1/minute over obscure modems*
Different interfaces	*1000's of languages in Europe alone*	*Switching was manual*	*//query (W5) inform?*
Material is intractable	*Scrolls and manu-scripts were about as random access as musical scores*	*No white pages*	*600 databases on Dialog ~1 Terabyte 140Gbyte at DJ 80GB card catalog at RLG*
Business model needed	*Centralized printing*	*Pay per minute*	*Not understood*

Navigation Techniques: Paper

- Alphabetical Listings (dictionary, Encyclopedia)
- Indices (back of the book and Readers Guide)
- Table of Contents (outlining)
- Citation index
- "Tree of Knowledge"
- Have you read any good books lately?

Navigation Techniques: Computers

- Hierarchical File Systems
- Unix "find" and "grep", Mac "find file"
- Boolean query systems (...within 5 words of...)
- Static Hypertext links (see also pointers)

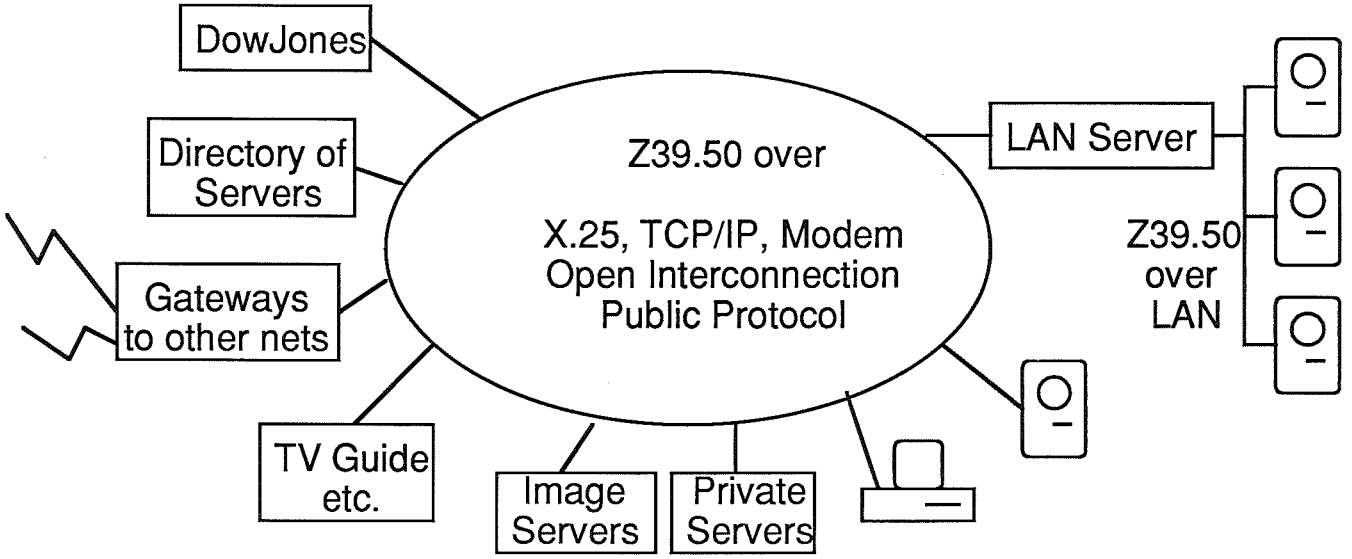
Navigation Techniques: WAIS

- English language questions and Relevance feedback
  * Iterative retrieval
  * Question-answer dialog
  * Similar to the Newspapers front page the: "continued on page 5"
  * Dynamic Hypertext Links

- 2 level search:
  * Directory of servers (server like any other)
  * Servers themselves

- Copy editors help select documents
  * Easy to "publish" opinions on documents

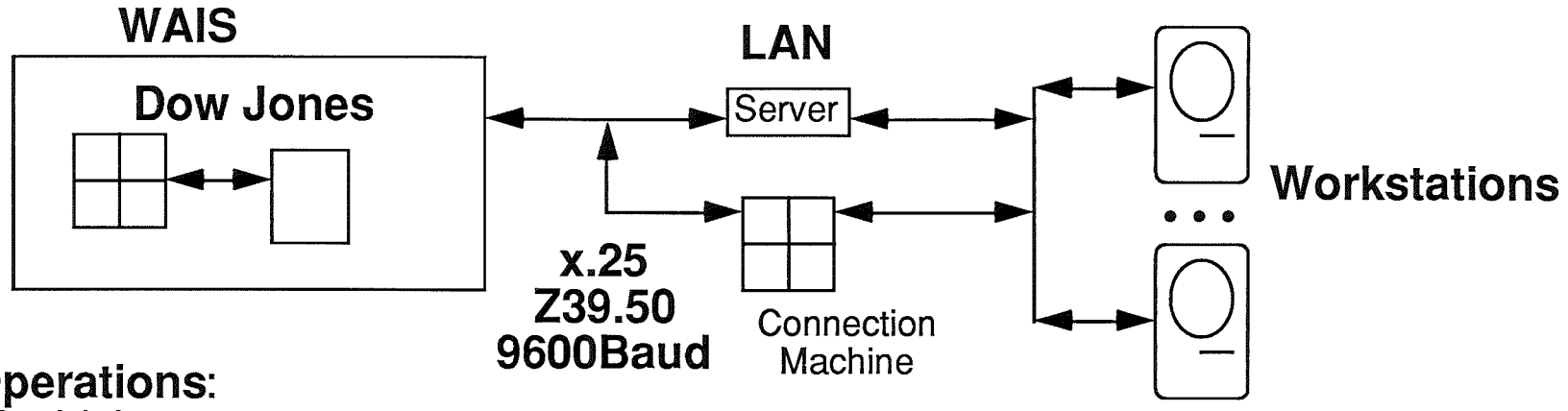
Wide Area Information Server Architecture



Users Needs:
- Selecting Servers
- Answering Questions
- Organizing Responses

Architecture Issues:
- Scalability
- Security
- Business model for servers
- Reliable Access

Demonstration System Structure



Operations:
Archiving
Queries
Retrieval
IR Type:
Broadcast
Query by Example
Databases:
Wall St Journal
Barron's
400 Business Mags

CM: Operations: Queries
IR Type:
enhanced relevance feedback
DBs: DowVision and
memo's, mail,
word processor files

Mac:
Operations:
Human Int
Retrieval
Queries
"Caching" Docs
User Profiles
IR Type:
Query by example
DBs:
Personal Text
Cached data

WAIS Clients

- Busy 24 hours a day finding information
- Ponder all indications of the preferences of its user
- Gossip with other clients about their discoveries
- Scours the world (within a budget) to find new sources

WAIS Protocol

- Based on Z39.50, bypass proprietary period
- Flexible
- Non Threatening for corporations

- Search: (words, doc_ids, databases) -> server
  returns list of: (headline, score, doc_id, types)'s
- Retrieval: (doc_id, type, start, end) -> server
  returns: bunch of bytes

- Doc_id: An ISBN for the Electronic Age
  ((orig_server, orig_database, orig_local_id)
  (dist_server, dist_database, dist_local_id)
- Server Description:
  (:ip-address, :database-name, :cost, :description)

Connection Machine Server

- 1-25GBytes (and getting bigger)
- Supports thousands of users
- Automatic Indexing
- Uses words and phrases in question to find appropriate documents
- First turn-key massively parallel application

TMC Internet Release

- CM product for TCP/IP (complete server)
- Example User interfaces for free (no support)
  Macintosh, Gnu Emacs, Xwindows
- Example unix server software to create servers
- Directory of Servers on the internet at least through '91
- 25 Servers now: Weather Maps, patents, Government programs, Risks-digest, usenet recipies, Lewis Carrol,...
- Anonymous FTP Think.com:/public/wais/*
  Mailing list: wais-discussion-request@think.com

Conclusion

- Electronic Publishing can fill niches now
- Companies are positioning themselves now (workstations, server, and info providers)
- Thinking Machines is the "Engine of the Information Industry"